Multi-diagnostic multi-model ensemble forecasts of aviation turbulence

Received: 16 February 2019 | Revised: 2 January 2020 | Accepted: 5 January 2020

DOI: 10.1002/met.1885

RESEARCH ARTICLE

Meteorological Applications RMetS

Multi-diagnostic multi-model ensemble forecasts of aviation turbulence

Luke N. Storer[1] | **Philip G. Gill**[2] | **Paul D. Williams**[1]

[1]Department of Meteorology, University of Reading, Reading, UK

[2]Met Office, Exeter, UK

Correspondence
Philip G. Gill, Met Office, Exeter, UK.
Email: philip.gill@metoffice.gov.uk

Funding information
Natural Environment Research Council, Grant/Award Number: NE/L002566; Royal Society, Grant/Award Number: UF130571

Abstract

Turbulence is one of the major weather hazards to aviation. Studies have shown that clear-air turbulence may well occur more frequently with future climate change. Currently the two World Area Forecast Centres use deterministic models to generate forecasts of turbulence. It has been shown that the use of multi-model ensembles can lead to more skilful turbulence forecasts. It has also been shown that the combination of turbulence diagnostics can also produce more skilful forecasts using deterministic models. This study puts the two approaches together to expand the range of diagnostics to include predictors of both convective and mountain wave turbulence, in addition to clear-air turbulence, using two ensemble model systems. Results from a 12 month global trial from September 2016 to August 2017 show the increased skill and economic value of including a wider range of diagnostics in a multi-diagnostic multi-model ensemble.

KEYWORDS

aviation, ensemble, forecast, multi-model, turbulence

1 | INTRODUCTION

Turbulence is a major hazard for the aviation industry, causing damage to aircraft and injury to passengers and flight crew. This can cost millions of dollars each year in compensation claims, lost working days for flight crew as well as aircraft maintenance time (Sharman and Lane, 2016; Gultepe *et al.*, 2019). The main type of turbulence that causes these injuries is clear-air turbulence (CAT). This is defined as high-altitude aircraft bumpiness in regions devoid of significant cloudiness and away from thunderstorm activity (Chambers, 1955). Because there is a lack of clouds, pilots are unable to foresee the turbulence, and without warning or preventative action taking place (such as putting the seatbelt sign on) objects including passengers can be tossed around the cabin. Recent research has also suggested that the frequency of CAT will increase in the future with climate change (Williams and Joshi, 2013; Storer *et al.*, 2017; Williams, 2017; Lee *et al.*, 2019), making the need for improved aviation turbulence forecasting even more important. There are three main sources of turbulence that impact aviation. The first is shear turbulence (Endlich, 1964; Atlas *et al.*, 1970) which is predominantly found around the fast flowing winds of the jet stream. The second is mountain wave turbulence (MWT) (Lilly, 1978) where gravity waves are produced which can ultimately lead to CAT (Storer *et al.*, 2019). The third is convectively induced turbulence (CIT) (Uccellini and Koch, 1987; Koch and Dorian, 1988)

where deep convection can initiate gravity waves, much like MWT, which can ultimately lead to CAT as well as in-cloud turbulence within the convective cloud. Pilots can visually see areas that may be at risk of turbulence within convective cloud from the use of on-board radar where the radar echoes give a measure of the intensity of precipitation. Although the radar reflectivity may be low for some convective cloud systems, the echo intensity does provide some indication of possible turbulence.

Currently the two World Area Forecast Centres (WAFCs) provide global gridded forecasts for aviation hazards including turbulence. These forecasts are derived from the output of deterministic models at a resolution of 1.25° with seven vertical levels (ICAO, 2016). However, previous studies have shown the improved performance in turbulence forecasting of probabilistic forecasts derived from ensemble models (Gill and Buchanan, 2014; Storer *et al.*, 2018a).

In Storer *et al.* (2018a) the idea of using a multi-model ensemble for aviation turbulence forecasting was introduced using only the Ellrod and Knapp (1992) Turbulence Index 1 (Ellrod TI1). Ellrod TI1 can only forecast shear turbulence and not MWT or CIT and therefore not all the turbulence events are forecasted. To address this issue, here multiple turbulence diagnostics and multiple ensemble forecasts are combined to create a multi-diagnostic multi-model ensemble. Storer *et al.* (2018a) showed that a single-diagnostic multi-model ensemble was able to increase forecast skill and relative economic value, although not at a statistically significant level. However, they also concluded that, by combining the two ensembles, a single authoritative forecast is created with an increased operational resilience.

Combining multiple turbulence diagnostics for an aviation turbulence forecast is not a new idea, because the Graphical Turbulence Guidance (GTG) system has been shown to improve forecast skill (Sharman *et al.*, 2006). In this study the newest generation Graphical Turbulence Guidance system 3 (GTG3) is used (Sharman and Pearson, 2017), not only forecasting shear turbulence but also including MWT (Sharman and Pearson, 2017). Studies combining two deterministic models (the National Oceanic and Atmospheric Administration Global Forecast System and the Met Office Unified Model) producing GTG diagnostics have been carried out to produce non-convective turbulence forecasts (Kim *et al.*, 2018). They created probabilistic forecasts by using probability density functions for each diagnostic rather than using the probabilities from an ensemble numerical weather prediction model.

This study, however, aims to bring together two different methods of turbulence forecasting (multi-diagnostic and multi-model ensemble) to see if there is any more skill that can be achieved that could be applied in the future to upgrade the current WAFC forecasts. To do this the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G) and the European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS) are combined. Three shear turbulence diagnostics are included which are the Ellrod TI1, the Brown Index (Brown, 1973) and the Richardson number (Ri). A MWT diagnostic is also included, which is the MWT12 from Sharman and Pearson (2017). A convectively induced turbulence predictor, which is the convective precipitation accumulation that was used by Gill and Buchanan (2014), is also included.

The method used in this study is the same as that of Storer *et al.* (2018a) which will be introduced in Section 2. In Section 3 the results including a single-diagnostic multi-model ensemble (Section 3.1), a multi-diagnostic multi-model ensemble (Section 3.2) and a reduced size multi-diagnostic multi-model ensemble (Section 3.3) are introduced. Section 4 presents conclusions and further work.

2 | METHODOLOGY

The method used for this multi-diagnostic multi-model ensemble is the same as that of Storer *et al.* (2018a). This trial uses a full year of ensemble forecast data between September 2016 and August 2017 from MOGREPS-G and EPS.

2.1 | Observations

The severity of turbulence experienced on board an aircraft is dependent on the type of aircraft. Therefore, it is important to use an aircraft-independent measure of turbulence as a smaller aircraft will experience more severe turbulence than a larger aircraft in the same volume of turbulent air. High-resolution automated aircraft data from the flight recorders on a fleet of Boeing 747 and 777 aircraft are used as the verification truth in the same way as for earlier verification studies (Gill, 2014; Storer *et al.*, 2018a). These observations are available at 4 s intervals and include, among others, the normal acceleration, airspeed and total aircraft mass at the time of observation. These data are used to calculate an aircraft-independent turbulence measure known as the derived equivalent vertical gust (DEVG), defined as:

$$\mathrm{DEVG} = \frac{Am\,|\,\Delta n\,|}{V} \tag{1}$$

where Δn is the peak modulus value of deviation of aircraft acceleration from 1g in units of g, m is the total

mass of the aircraft (metric tonnes), V is the calibrated airspeed at the time of the observation (knots) and A is an aircraft-specific parameter which varies with flight conditions and is defined as:

$$A = \bar{A} + c_4(\bar{A} - c_5)\left(\frac{m}{\bar{m}-1}\right) \tag{2}$$

$$\bar{A} = c_1 + \frac{c_2}{c_3 + H} \tag{3}$$

where H is the altitude (thousands of feet), $\bar{m}$ is the reference mass of the aircraft (metric tonnes) and the aircraft-dependent parameters c_1 *to* c_5 depend on the aircraft's flight profile as outlined in Truscott (2000). The value of 4.5 $m{\cdot}s^{-1}$ for DEVG is taken as the threshold for moderate or greater turbulence. The distribution of aircraft data is weighted towards the Northern Hemisphere and in particular the North Atlantic, Europe and North America; however, there is regular coverage for parts of Africa, South America, Asia and Australia as illustrated in Figure 1, showing a sample plot of aircraft data for May 2017.

2.2 | Turbulence predictors

The non-convective turbulence diagnostics used in this trial are taken from GTG3: the Ellrod TI1, the Brown index (Brown, 1973), a MWT predictor (MWT12 from Sharman and Pearson, 2017) and the Richardson number (Ri) are used. The convective precipitation accumulation, used as a CIT predictor, is formed from combining the convective snow and convective rain accumulations which are taken directly from numerical weather prediction models.

The Ellrod TI1 turbulence diagnostic is defined as the product of the deformation (DEF) and vertical wind shear (VWS):

$$\mathrm{TI1} = \mathrm{DEF} \times \mathrm{VWS} = \left\{\frac{\partial u}{\partial x} - \frac{\partial v^2}{\partial y} + \frac{\partial v}{\partial x} + \frac{\partial u^2}{\partial y}\right\}^{\frac{1}{2}} \times \left\{\frac{\partial u^2}{\partial z} + \frac{\partial v^2}{\partial z}\right\}^{\frac{1}{2}} \tag{4}$$

where u is the horizontal wind velocity in the east–west direction, v is the horizontal wind velocity in the north–

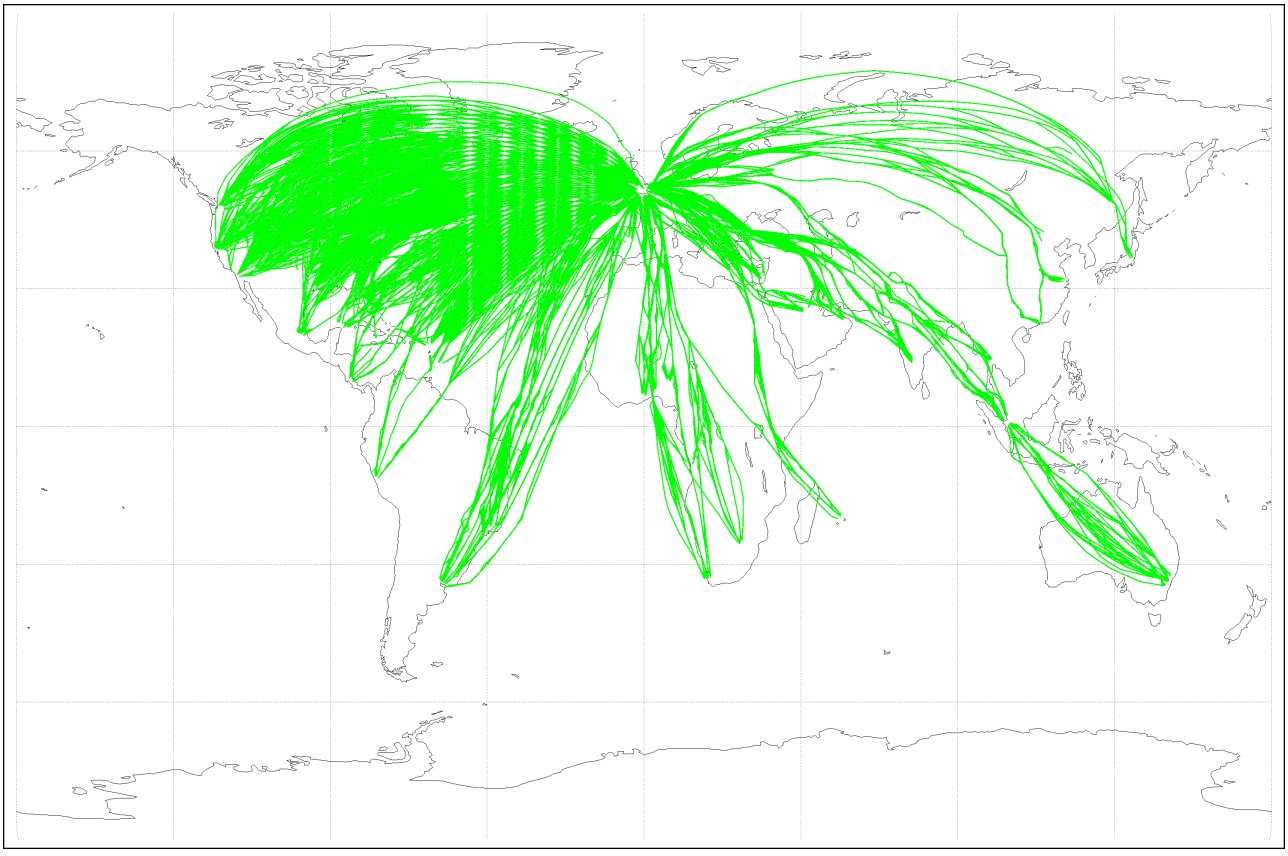

FIGURE 1 Map showing an example of the distribution of verifying aircraft observations for May 2017

south direction, x is distance in the east–west direction, y is distance in the north–south direction and z is distance in the vertical.

The Brown index was used as one of the turbulence predictors in Gill (2014) and is useful because it includes absolute vorticity, shearing and stretching deformation; it is defined as:

$$\text{Brown} = \left\{ 0.3 \left(\frac{\partial v}{\partial x} - \frac{\partial u}{\partial y} + f \right)^2 + \left(\frac{\partial v}{\partial x} + \frac{\partial u}{\partial y} \right)^2 + \left(\frac{\partial u}{\partial x} - \frac{\partial v}{\partial y} \right)^2 \right\}^{\frac{1}{2}} \tag{5}$$

where u, v, x and y are as defined in Equation (4) and f is the coriolis frequency.

The MWT predictor used in this project is MWT12 from Sharman and Pearson (2017) and is the one that performed best in their trial over the United States; it is defined as:

$$\text{MWT} = d_s \times |\text{TEMPG}| \tag{6}$$

where d_s is a near surface diagnostic (low-level wind speed perpendicular to the ridgeline) and TEMPG is the horizontal temperature gradient.

The Richardson number is defined as:

$$\text{Ri} = \frac{N^2}{(\partial U/\partial z)^2} = \frac{(g/\theta)(\partial \theta/\partial z)}{(\partial U/\partial z)^2} \tag{7}$$

where $N^2$is the Brunt–Väisälä frequency squared, U is horizontal wind speed, z is altitude, g is gravitational acceleration and θ is potential temperature.

Convective precipitation is used as a proxy for convectively induced turbulence following on from earlier studies (Gill and Stirling, 2013). The convective diagnostic from the model is on a higher resolution grid, so an interpolation routine from the Met Office Iris library (Iris, 2015) is used to regrid the diagnostic linearly so that it has the same horizontal resolution as the other turbulence diagnostics.

2.3 | Probabilistic forecasts

At the time of this trial, MOGREPS-G had 12 members and 70 levels with a horizontal resolution of 33 km and EPS had 51 members (referred to in this paper as EPS51) and 91 levels with a horizontal resolution of 20 km. The forecasts used in this project are output to a horizontal resolution of 1° and 26 vertical levels (although only six are useful for aviation at cruise levels), rather than their native grid (which was used in Storer *et al.*, 2018a). Forecasts output from 0000 UTC for $T + 24$, $T + 27$, $T + 30$, $T + 33$ hr are used. It is important to keep this in mind when analysing the results as the focus will only be on half the day from 2230 UTC to 1030 UTC, and this may not give a complete picture of performance.

To create a probabilistic turbulence forecast, thresholds are set for each of the turbulence diagnostics and any time an ensemble member exceeds that threshold it is classed as a turbulence forecast. The proportion of ensemble members gives an estimate of the probability of exceeding each turbulence threshold. The thresholds are chosen to divide the distribution of forecast values approximately equally. The thresholds used for each diagnostic and ensemble are shown in Table 1 and they are the same for MOGREPS-G and EPS. The convective precipitation accumulation threshold is numerically different, but this is because the thresholds needed are much higher for MOGREPS-G than for EPS because the units are different. MOGREPS-G has units of kg·m^{-2} whereas EPS has units of metres. The two units are related, however, because 1 kg·m^{-2} is equivalent to 1 mm of precipitation (which is the same as 0.001 m). This therefore means that the thresholds are actually the same but the units are different.

TABLE 1 The five turbulence thresholds used for each of the turbulence diagnostics in this study

Diagnostic	Units	Ensemble	Thr1	Thr2	Thr3	Thr4	Thr5
Ellrod T1 × 10^{-7}	s^{-2}	MOGREPS-G and EPS	3	5	8	11	20
Brown × 10^{-5}	s^{-1}	MOGREPS-G and EPS	12	15	20	25	30
MWT × 10^{-1}	k·s^{-1}	MOGREPS-G and EPS	1	5	10	15	20
Richardson × 10^{-2}		MOGREPS-G and EPS	5	10	20	50	100
Convection × 10^{-2}	×10^3 kg·m^{-2}	MOGREPS-G	1	5	10	15	20
	m	EPS					

Ellrod T1, Ellrod and Knapp (1992) Turbulence Index 1; EPS, European Centre for Medium-Range Weather Forecasts Ensemble Prediction System; MOGREPS-G, Met Office Global and Regional Ensemble Prediction System; MWT, mountain wave turbulence; Thr, threshold.

2.4 | Verification

The verification method is also the same as Storer *et al.* (2018a). For each flight, the maximum observation within each 10 min flight segment is calculated. This is compared to the corresponding forecast by using bilinear interpolation on the forecast grid to the location of the aircraft at the point of the maximum observed value within the flight segment. The model with closest validity time to the observation was used. This method gives a time window of ±1.5 hr around each model validity time, allowing all observations to be used. For each probability threshold a contingency table is created by applying the threshold on the observation for moderate or greater turbulence.

From these results a relative operating characteristic (ROC) curve can be plotted which shows forecast skill by plotting the hit rate against the probability of false detection (Jolliffe and Stephenson, 2012; Gill, 2016), which are defined as:

$$\text{hit rate}\,(H) = \frac{A}{A+C} \tag{8}$$

$$\text{probability of false detection}\,(\text{POFD}) = \frac{B}{B+D} \tag{9}$$

where A is a hit, B is a false alarm, C is a miss and D is a correct rejection. The more skilful the forecast, the higher the area under the curve (AUC) will be (over a baseline value of 0.5 representing zero skill), and this shows a forecast that has found a good balance between forecasting as many hits as possible while minimizing the number of false alarms.

The forecast skill shown by a ROC plot is not the only way to show how useful a forecast is. The relative economic value (Richardson, 2000; Jolliffe and Stephenson, 2012) shows how valuable the forecast is for a given cost/loss ratio which will be user specific. If the forecast is more valuable for all cost/loss ratios, known as sufficiency (Ehrendorfer and Murphy, 1988), any user will benefit from this model.

Forecast reliability is also a way of understanding how well a forecast model performs (Jolliffe and Stephenson, 2012; Gill, 2016). By plotting the forecasted probability and the actual observed frequency, it can indicate if the turbulence events are being over- or under-forecasted.

2.5 | Combining diagnostics

The probabilistic forecast for each of the turbulence diagnostics is created first, before combining them. To do this the same method as shown in Storer *et al.* (2018a) is followed by creating a single-diagnostic single-model predictor but this time for each of the five turbulence diagnostics.

The multi-diagnostic multi-model ensemble predictor can be created in two ways, first by combining the probabilistic forecasts from all diagnostics at each threshold and from both ensemble models equally to create a simple super ensemble.

The second method is to combine the probabilistic forecasts by optimizing the weighting for each diagnostic and threshold and from each model used. The optimization works by first prescribing a set of initial weightings for each diagnostic and each threshold, and then running an iterative scheme to change the weightings in 0.1 increments to optimize the weightings by maximizing the AUC. If the greatest AUC from this process exceeds the current AUC then the predictor becomes part of the weighted sum and the process moves on to the next predictor. The process repeats until the AUC can no longer be increased, and in some cases this may result in some diagnostics having zero weight for some thresholds but not others. The resulting set of weights is then used to create an optimized predictor. This optimization method was also used by Gill and Buchanan (2014). It is important to note that this is a trial and error method (for the initial weightings) and therefore the AUC obtained will not be the best the models can achieve.

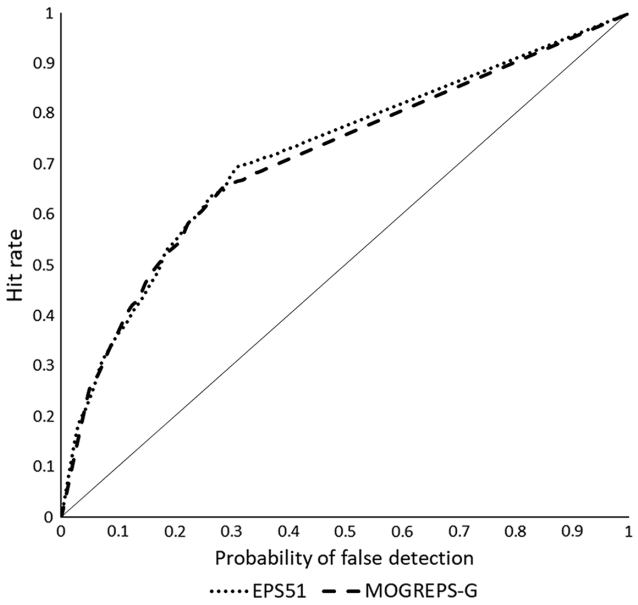

FIGURE 2 Receiver operating characteristic plot of global turbulence forecasts for threshold 1 (3×10^{-7} s^{-2}) of the Ellrod and Knapp (1992) Turbulence Index 1 diagnostic for the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G) and the European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (EPS51). The data used have a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017

TABLE 2 The area under the curve for five thresholds for each of the five turbulence diagnostics for EPS51, MOGREPS-G, the combined multi-model ensemble 95% lower confidence interval, the combined multi-model ensemble and the combined multi-model ensemble 95% upper confidence interval

	Thr1	Thr2	Thr3	Thr4	Thr5
Ellrod TI1					
EPS51	0.7197	0.7032	0.6486	0.5978	0.5162
MOGREPS-G	0.7111	0.6905	0.6272	0.5854	0.5162
Multi-model 95% lower CI	0.7013	0.6973	0.6378	0.5923	0.5126
Multi-model	0.7244	0.7197	0.6601	0.6104	0.5210
Multi-model 95% upper CI	0.7477	0.7425	0.6837	0.6298	0.5306
Brown index					
EPS51	0.6189	0.5781	0.5311	0.5170	0.5035
MOGREPS-G	0.5912	0.5595	0.5276	0.5125	0.5056
Multi-model 95% lower CI	0.6026	0.5643	0.5199	0.5103	0.5013
Multi-model	0.6244	0.5858	0.5328	0.5186	0.5054
Multi-model 95% upper CI	0.6476	0.6065	0.5454	0.5280	0.5105
MWT					
EPS51	0.5293	0.5038	0.5000	0.5000	0.5000
MOGREPS-G	0.5298	0.5055	0.5020	0.5000	0.5000
Multi-model 95% lower CI	0.5182	0.5015	0.4999	0.5000	0.5000
Multi-model	0.5302	0.5064	0.5019	0.5000	0.5000
Multi-model 95% upper CI	0.5438	0.5121	0.5051	0.5000	0.5000
Richardson number					
EPS51	0.7694	0.7703	0.6833	0.5210	0.5040
MOGREPS-G	0.7694	0.7703	0.6730	0.5231	0.5040
Multi-model 95% lower CI	0.7596	0.7650	0.6813	0.5219	0.5009
Multi-model	0.7773	0.7864	0.7036	0.5326	0.5040
Multi-model 95% upper CI	0.7951	0.8060	0.7255	0.5445	0.5086
Convective precipitation accumulation					
EPS51	0.7142	0.7216	0.6994	0.6812	0.6480
MOGREPS-G	0.7126	0.7059	0.6855	0.6720	0.6611
Multi-model 95% lower CI	0.6968	0.6987	0.6833	0.6686	0.6534
Multi-model	0.7195	0.7222	0.7076	0.6924	0.6767
Multi-model 95% upper CI	0.7439	0.7461	0.7303	0.7154	0.6997

CI, confidence interval; Ellrod T1, Ellrod and Knapp (1992) Turbulence Index 1; EPS, European Centre for Medium-Range Weather Forecasts Ensemble Prediction System; MOGREPS-G, Met Office Global and Regional Ensemble Prediction System; MWT, mountain wave turbulence; Thr, threshold.
The data used have a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017.

3 | RESULTS

3.1 | Single-diagnostic ensembles

Figure 2 is an example ROC plot of the Ellrod TI1 turbulence diagnostic for EPS51 and MOGREPS-G using the first and therefore lowest turbulence threshold. The MOGREPS-G curve is slightly below the EPS51 curve, particularly at the higher values of the POFD. This will lead to a slightly higher AUC for EPS51 than MOGREPS-G for the single-diagnostic single-model predictor.

The MOGREPS-G, EPS51 and combined equal weighted single-diagnostic multi-model ensemble AUC is shown in Table 2 for each of the turbulence diagnostics and all five thresholds. EPS51 has a higher AUC for most of the diagnostics and thresholds. It is interesting that the two highest thresholds for MWT have no skill, as the threshold is too high to capture any turbulence events.

This is not a surprise because the MWT predictor will only be forecasting events over and around mountains. The number of observations in these areas and therefore the number of events that could be forecasted are much lower than the other turbulence diagnostics that forecast, for example, around the jet stream. The multi-model ensemble is slightly more skilful than either EPS51 or MOGREPS-G for each predictor. This follows on from Storer *et al.* (2018a) and shows that having a multi-model ensemble is more skilful than a single-model ensemble. However, as in Storer *et al.* (2018a), statistical significance with the 95% confidence interval cannot be shown. Both the upper and lower confidence interval bounds for the multi-model ensemble are shown in Table 2 and there are only six occasions where the multi-model ensemble is significantly more skilful than MOGREPS-G, only one occasion for EPS51 and there are no occasions when it is significantly higher than both.

To illustrate this, the two thresholds with the highest AUC for each of the diagnostics (which is thresholds 1 and 2 for all diagnostics) are plotted in Figure 3, comparing MOGREPS-G, EPS51 and the combined multi-model ensemble. The 95% confidence intervals for the multi-model ensemble are also included. It is seen that for the most part EPS51 is more skilful than MOGREPS-G and there are only three occasions where the multi-model ensemble is significantly more skilful than either of the single-model ensembles. Figure 3 also shows how well each of the different diagnostics performs, with the Richardson number producing the highest AUC and Ellrod TI1 and convective precipitation accumulation just behind. The Brown index and MWT perform worse with much lower AUC; however, including them may still have some benefit in a multi-diagnostic ensemble as they may be forecasting more extreme events that the other predictors could be missing. This is especially true of the MWT predictor as none of the other predictors is targeting MWT and therefore, without it, these events will be missed.

3.2 | Multi-diagnostic ensembles

In this study both the equally weighted and optimized multi-diagnostic predictor for MOGREPS-G, EPS51 and a multi-diagnostic multi-model ensemble are created (Figure 4). The weightings used in the multi-diagnostic multi-model ensemble are shown in Table 3. Some of the diagnostics have more weight (e.g. Richardson number and convective precipitation accumulation) and for some diagnostics one model has no weight (e.g. EPS51 for the Brown index). The ROC plot for the optimized multi-diagnostic predictors is shown in Figure 5 and the areas under the ROC curves, together with confidence intervals, are given in Table 4 and Figure 6. The first thing to note from the ROC plot in Figure 5 is that it has a much higher AUC than the single-diagnostic predictors shown in Figure 2. This shows that using multiple thresholds and diagnostics can increase the forecast skill. It is also interesting that MOGREPS-G performs better than EPS51 and at very low values of the POFD MOGREPS-G is slightly more skilful than the multi-model ensemble. This is a very interesting result and one that was not expected. In Section 3.1 it was found that EPS51 has a consistently higher AUC for almost every diagnostic, yet when combining diagnostics it is seen that the MOGREPS-G is more skilful than EPS51.

For a single-diagnostic ensemble to be skilful, the forecast spread needs to be large enough to capture as

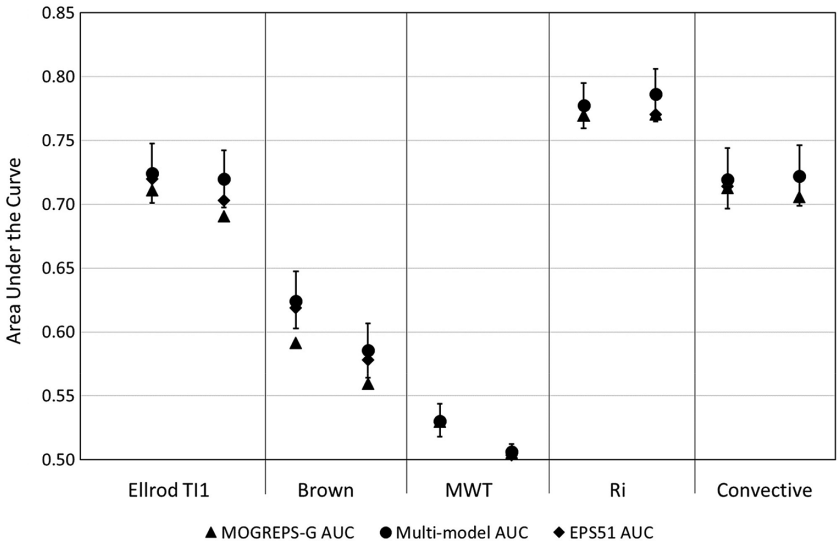

FIGURE 3 Plot showing the area under the curve for thresholds 1 and 2 for five turbulence diagnostics from the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G) (triangles), the European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (EPS51) (diamonds) and the combined multi-model ensemble (circles). The combined single-diagnostic multi-model ensemble has error bars showing the 95% confidence interval. The data used have a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017

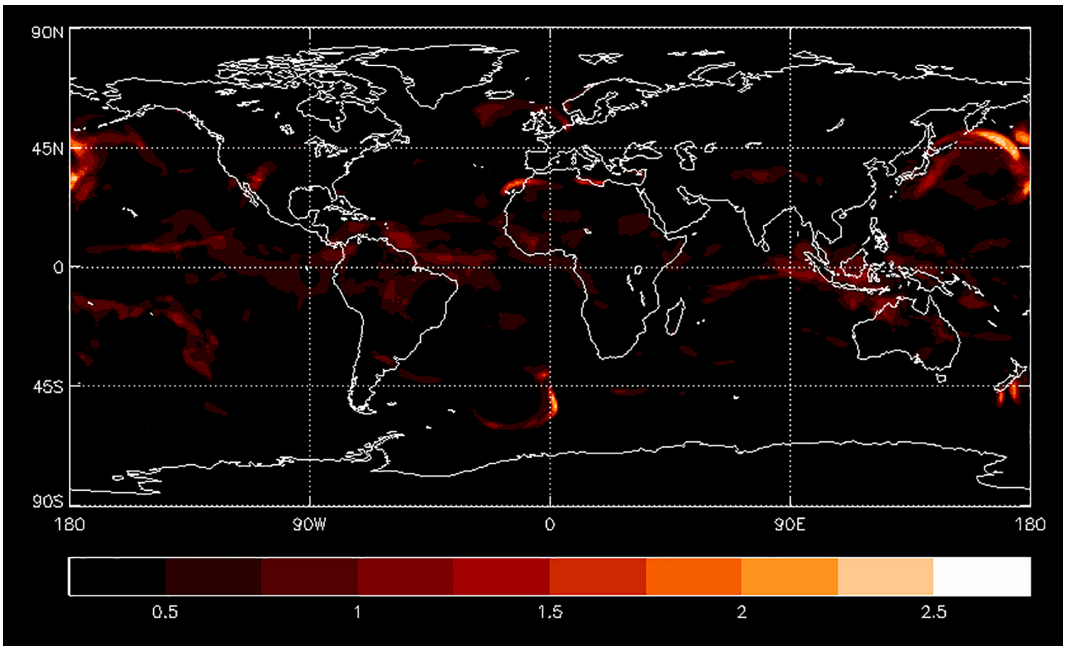

FIGURE 4 Example of calibrated probabilistic multi-model ensemble multi-predictor $T + 24$ forecast for the probability of moderate or greater turbulence for December 22, 2016 at 00Z

TABLE 3 The weightings used with MOGREPS-G and EPS51 to create the optimized multi-diagnostic multi-model ensemble

Diagnostic	Ensemble	Thr1	Thr2	Thr3	Thr4	Thr5
Ellrod T1	MOGREPS-G	0.1	0.1	0.1	0.4	0.1
	EPS51	0.0	0.1	0.1	1.3	0.1
Brown	MOGREPS-G	0.0	0.1	0.1	0.0	0.9
	EPS51	0.0	0.0	0.0	0.0	0.0
MWT	MOGREPS-G	0.1	0.0	0.0	0.0	0.0
	EPS51	0.3	0.0	0.0	0.0	0.0
Richardson	MOGREPS-G	0.1	0.2	0.2	0.9	0.1
	EPS51	0.1	0.1	0.1	0.3	0.1
Convection	MOGREPS-G	0.2	0.1	0.1	0.1	0.2
	EPS51	0.0	0.1	0.2	0.1	0.1

Ellrod T1, Ellrod and Knapp (1992) Turbulence Index 1; EPS, European Centre for Medium-Range Weather Forecasts Ensemble Prediction System; MOGREPS-G, Met Office Global and Regional Ensemble Prediction System; MWT, mountain wave turbulence; Thr, threshold.

many turbulence events as possible (hits) but also to avoid too many false alarms. EPS51 achieves this more successfully than MOGREPS-G, which is why it has a higher AUC for almost all the diagnostics (Figure 3). However, if each of the diagnostics forecasts a turbulence event slightly differently, when combining ensembles the area of forecasted turbulence will be increased. This can lead to more hits but there could also be more false alarms. The larger the ensemble spread of the individual diagnostics, the greater the ensemble spread of the multi-diagnostic ensemble. It appears in this case that, although the spread for EPS51 works well for the individual diagnostics, when combining them the number of false alarms starts to outweigh the number of hits and the resultant forecast skill is reduced. MOGREPS-G, however, has less forecast spread for the individual diagnostics but then, when combining them, the forecast spread increases just the right amount and the forecast skill is then higher than EPS51.

Table 4 shows the AUC for the combined equal and combined optimized single-model and multi-model ensembles. The multi-model ensemble also has the 95% upper and lower confidence intervals and again it is not significantly better than either of the two single-model

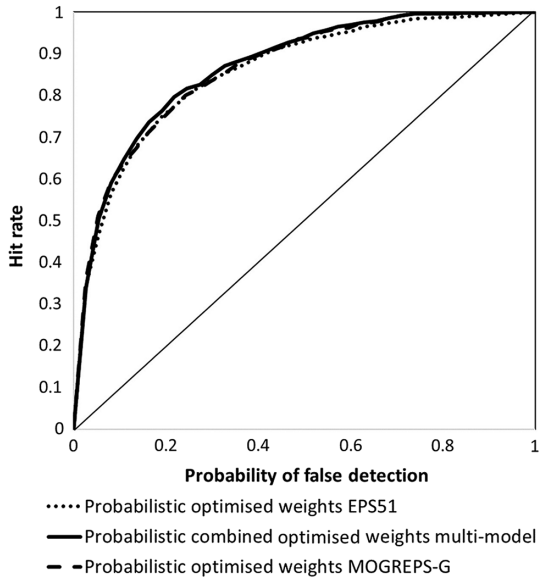

FIGURE 5 Receiver operating characteristic (ROC) plot of global turbulence forecasts for the multi-diagnostic Met Office Global and Regional Ensemble Prediction System (MOGREPS-G) (dashed), the multi-diagnostic European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (EPS51) (dotted) and the combined multi-diagnostic multi-model ensemble (solid line). Five turbulence thresholds for each turbulence diagnostic are optimally combined to maximize the area under the ROC curve and use a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017

ensembles. Figure 6 shows the results in Table 4 as a bar chart and it can clearly be seen that, for both the combined equal and combined optimized ensemble, MOGREPS-G is more skilful. The combined optimized ensemble is also more skilful than the combined equal ensemble and shows the benefit of taking time to optimize the thresholds and diagnostics used.

Figure 7 is a value plot of the optimized multi-diagnostic ensemble for MOGREPS-G, EPS51 and the multi-model ensemble. It is seen that it is MOGREPS-G that has more value for all cost/loss ratios than EPS51 which is the opposite of what was found by Storer *et al.* (2018a) using a single diagnostic. This follows on from the ROC plots where MOGREPS-G outperformed EPS51 and again suggests that, when combining the diagnostics, the spread from EPS51 is too large and therefore the balance between hits and false alarms is not quite optimized. What is also seen is that the combined multi-model ensemble is not more valuable for all cost/loss ratios. This suggests that, for some consumers, it might be more beneficial to just have MOGREPS-G and not include EPS51 at all. The differences in value are small and may not be significantly different; however, the multi-model ensemble would add operational resilience and the multi-model

TABLE 4 Area under the curve for the combined equal weighting multi-diagnostic ensemble and combined optimized multi-diagnostic ensemble for EPS51, MOGREPSG, combined multi-model ensemble 95% lower confidence interval, combined multi-model ensemble and combined multi-model ensemble 95% upper confidence interval

	Combined equal	Combined optimized
EPS51	0.8498	0.8555
MOGREPS-G	0.8564	0.8630
Multi-model 95% lower CI	0.8442	0.8530
Multi-model	0.8590	0.8679
Multi-model 95% upper CI	0.8745	0.8829

CI, confidence interval; EPS, European Centre for Medium-Range Weather Forecasts Ensemble Prediction System; MOGREPS-G, Met Office Global and Regional Ensemble Prediction System.
The data used have a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017.

ensemble is almost as valuable as MOGREPS-G and would therefore be the better option.

Forecast reliability is shown in Figure 8 and a calibration constant has been applied to the forecast probability to improve the reliability by reducing the over-forecasting without calibration. This constant is 1/50 for MOGREPS-G, 1/40 for EPS51 and 1/130 for the multi-model ensemble. What is found is that there is a reasonable match between the forecast probability and observed frequency; however, the probabilities are limited to low values following calibration. Although the probabilities following calibration are low, compared to the background frequency of moderate or greater turbulence (0.096%), the forecasts can still be a useful warning of higher than average turbulence probabilities. The background frequency of turbulence can be obtained from the verification by looking at the number of turbulence events (hits and misses) as a proportion of the total sample. It could be possible to apply a more sophisticated nonlinear calibration; however, for this study it has been kept simple with the linear constant.

3.3 | Reduced size multi-diagnostic ensemble

As in Storer *et al.* (2018a), it is important to understand how the two ensembles compare to each other with the same ensemble size. The ECMWF ensemble has 51 members, consisting of one control member and 50 equally likely perturbed members. Each consecutive member has a pairwise anti-symmetric perturbation (Owens and Hewson, 2018). This can therefore be reduced by

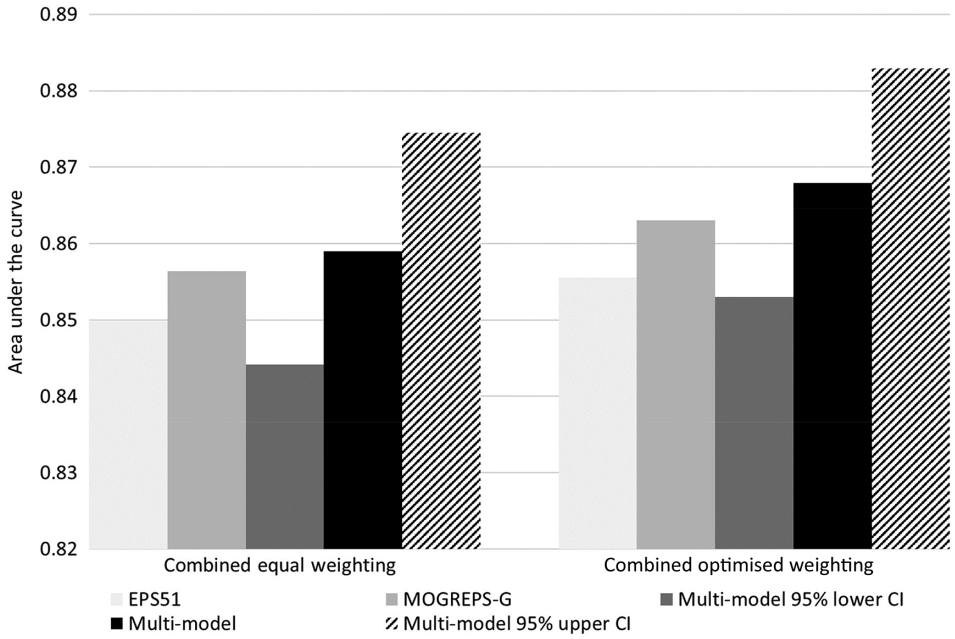

FIGURE 6 Bar chart showing the area under the curve (AUC) for the multi-diagnostic European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (EPS51) (light grey), the multi-diagnostic Met Office Global and Regional Ensemble Prediction System (MOGREPS-G) (mid grey), the multi-diagnostic multi-model ensemble 95% lower confidence interval (dark grey), the combined multi-diagnostic multi-model ensemble (black) and the combined multi-diagnostic multi-model ensemble 95% upper confidence interval (diagonal stripes). For the bar chart on the left the five turbulence thresholds for each turbulence diagnostic are combined equally and on the right the five turbulence thresholds for each turbulence diagnostic are optimally combined to maximize the AUC. The data used have a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017

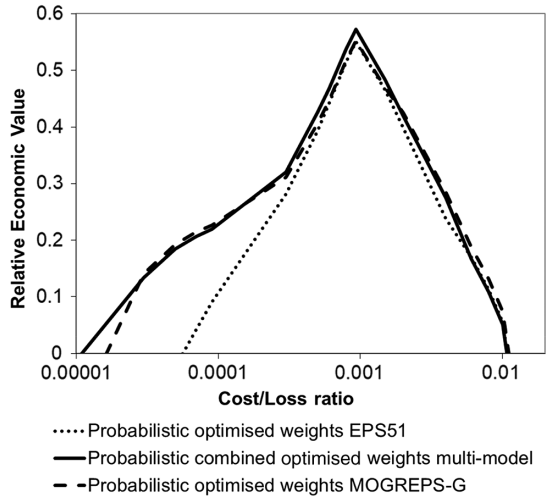

FIGURE 7 Value plot with a log scale x-axis of the global turbulence forecasts showing the forecast skill for the multi-diagnostic Met Office Global and Regional Ensemble Prediction System (MOGREPS-G), the multi-diagnostic European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (EPS51) and the combined multi-diagnostic multi-model ensemble. Five turbulence thresholds for each turbulence diagnostic are optimally combined to maximize the area under the receiver operating characteristic curve and use a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017

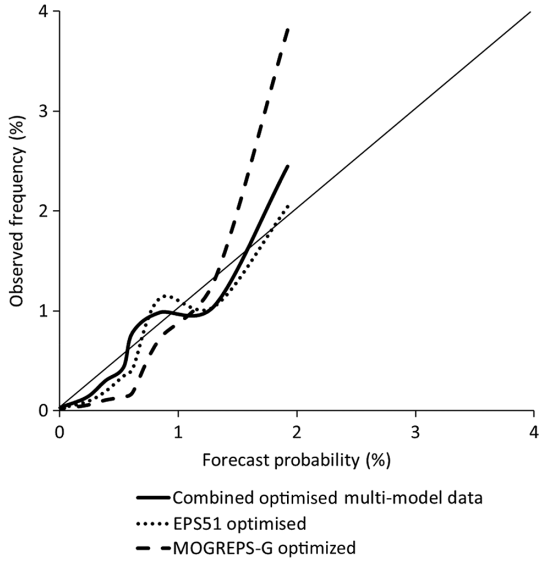

FIGURE 8 Reliability diagram for the multi-diagnostic Met Office Global and Regional Ensemble Prediction System (MOGREPS-G), the multi-diagnostic European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (EPS51) and the combined multi-diagnostic multi-model ensemble. Five turbulence thresholds for each turbulence diagnostic are optimally combined to maximize the area under the receiver operating characteristic curve and use a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017

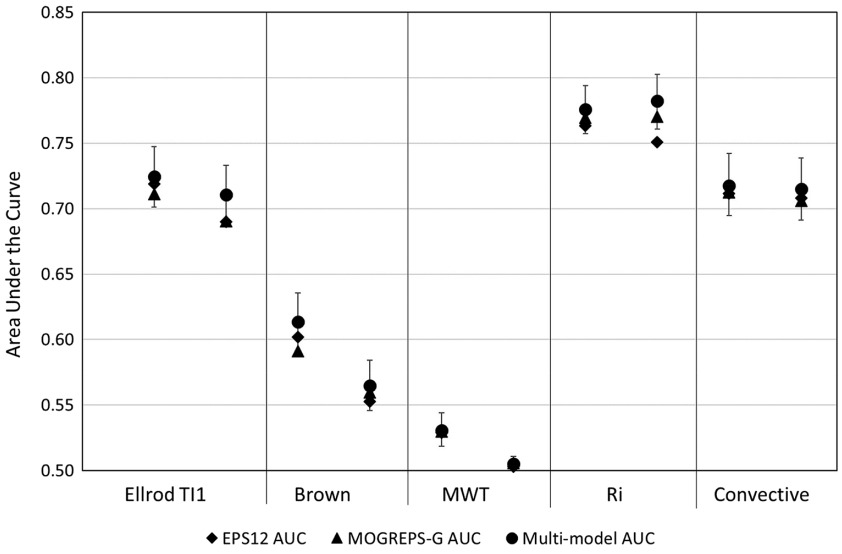

FIGURE 9 Plot showing the area under the curve (AUC) for thresholds 1 and 2 for five turbulence diagnostics from the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G) (triangles), the European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (EPS12) (diamonds) and the combined multi-model ensemble (circles). The combined multi-model ensemble has error bars showing the 95% confidence interval. The data used have a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017

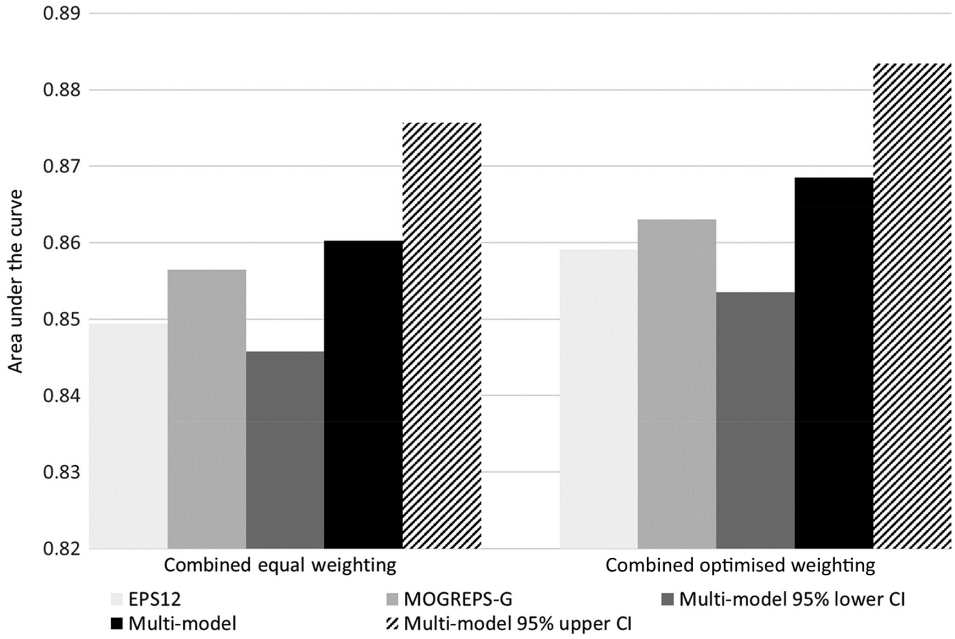

FIGURE 10 Bar chart showing the area under the curve (AUC) for the multi-diagnostic European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (EPS12) (light grey), the multi-diagnostic Met Office Global and Regional Ensemble Prediction System (MOGREPS-G) (mid grey), the combined multi-diagnostic multi-model ensemble 95% lower confidence interval (dark grey), the combined multi-diagnostic multi-model ensemble (black) and the combined multi-diagnostic multi-model ensemble 95% upper confidence interval (diagonal stripes). For the bar chart on the left the five turbulence thresholds for each turbulence diagnostic are combined equally and on the right the five turbulence thresholds for each turbulence diagnostic are optimally combined to maximize the AUC. The data used have a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017

selecting the first 12 members and producing a bias-free subsample, referred to from now on as EPS12, to compare directly with the 12-member MOGREPS-G. A similar approach to investigate ensemble size and ensemble skill was taken by Buizza and Palmer (1998). The AUC for the first two thresholds for each turbulence diagnostic is shown in Figure 9. The circles are the multi-model ensemble with the 95% confidence intervals displayed; MOGREPS-G (triangles) and EPS12 (diamonds) are also shown. As in Figure 3 the multi-model ensemble is more skilful than both of the single-model ensembles and for the most part EPS12 is slightly more skilful than

Diagnostic	Ensemble	Thr1	Thr2	Thr3	Thr4	Thr5
Ellrod T1	MOGREPS-G	0.1	0.1	0.1	0.8	0.2
	EPS12	0.0	0.1	0.1	0.5	0.1
Brown	MOGREPS-G	0.0	0.0	0.2	0.0	0.0
	EPS12	0.0	0.0	0.0	0.0	0.0
MWT	MOGREPS-G	0.1	0.0	0.0	0.0	0.0
	EPS12	0.4	0.0	0.0	0.0	0.0
Richardson	MOGREPS-G	0.1	0.1	0.1	1.5	0.1
	EPS12	0.0	0.1	0.2	0.1	0.1
Convection	MOGREPS-G	0.1	0.1	0.1	0.1	0.1
	EPS12	0.0	0.1	0.2	0.1	0.1

TABLE 5 The weightings used with MOGREPS-G and EPS12 to create the optimized multi-diagnostic multi-model ensemble

Ellrod T1, Ellrod and Knapp (1992) Turbulence Index 1; EPS, European Centre for Medium-Range Weather Forecasts Ensemble Prediction System; MOGREPS-G, Met Office Global and Regional Ensemble Prediction System; MWT, mountain wave turbulence; Thr, threshold.

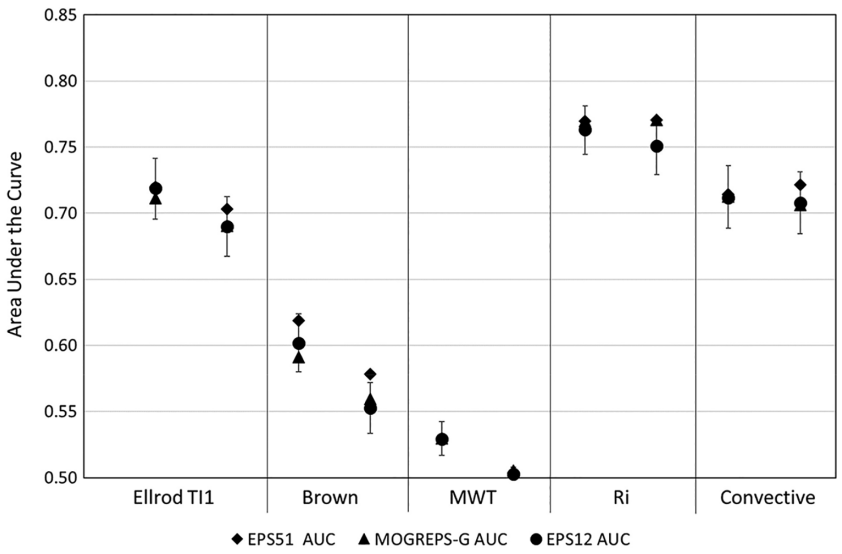

FIGURE 11 Plot showing the area under the curve (AUC) for thresholds 1 and 2 for five turbulence diagnostics from the Met Office Global and Regional Ensemble Prediction System (MOGREPS-G) (triangles), the European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (EPS51) (diamonds) and the EPS12 (circles). EPS12 has error bars showing the 95% confidence interval. The data used have a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017

MOGREPS-G. It is interesting to note that EPS12 seems to be slightly less skilful than EPS51, which is what was found by Storer *et al.* (2018a) but here it is consistent across all turbulence diagnostics.

Figure 10 is a bar chart showing the combined multi-diagnostic predictor for EPS12, MOGREPS-G and the combined multi-model ensemble with its upper and lower 95% confidence intervals. The weightings used in the optimized multi-diagnostic multi-model ensemble are shown in Table 5. It is found that the multi-model ensemble with EPS12 is more skilful than the multi-model ensemble that uses EPS51 (from Figure 6). This agrees with Section 3.2 that a smaller ensemble spread for each individual diagnostic will give an overall better performance when combined in a multi-diagnostic ensemble. This helps to explain why MOGREPS-G is more skilful than EPS51.

To investigate this further, Figure 11 is a plot showing the AUC for EPS12 with its 95% confidence interval displayed, EPS51 and MOGREPS-G for the individual diagnostics. As expected, EPS51 is more skilful than EPS12 and in some cases is significantly better. MOGREPS-G has a similar forecast skill to EPS12 and therefore the reduced number of members results in a lower AUC for the individual diagnostics. However, Figure 12 shows that the equal combined multi-diagnostic EPS51 is fractionally more skilful than EPS12, but when optimized the EPS12 is slightly more skilful. It is also interesting that MOGREPS-G is more skilful than either the EPS12 or EPS51 multi-diagnostic ensembles.

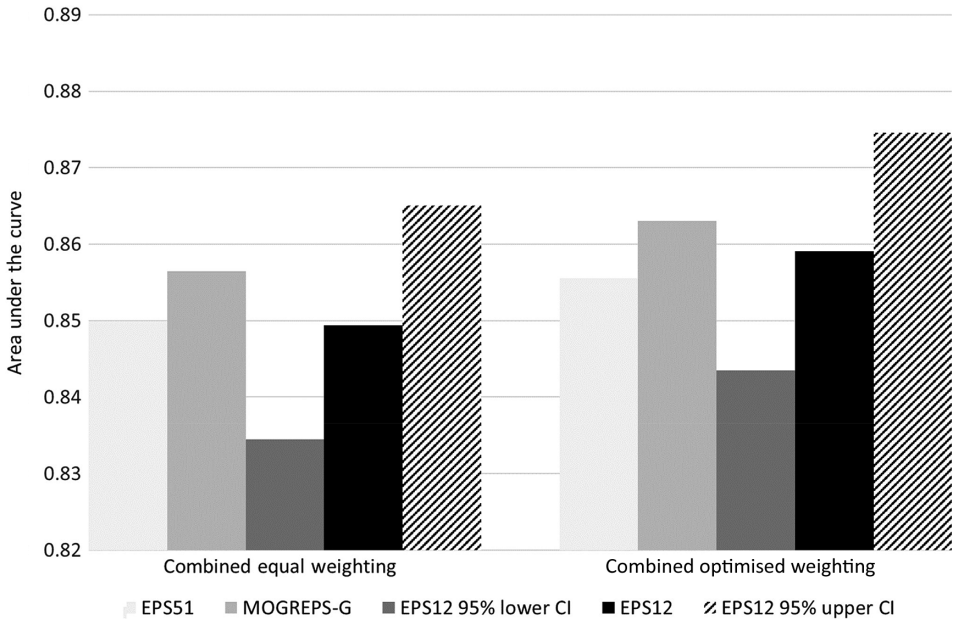

FIGURE 12 Bar chart showing the area under the curve (AUC) for the multi-diagnostic European Centre for Medium-Range Weather Forecasts Ensemble Prediction System (EPS51), the multi-diagnostic Met Office Global and Regional Ensemble Prediction System (MOGREPS-G), the multi-diagnostic EPS12 95% lower confidence interval, the multi-diagnostic EPS12 and the multi-diagnostic EPS12 95% upper confidence interval. For the bar chart on the left the five turbulence thresholds for each turbulence diagnostic are combined equally and on the right the five turbulence thresholds for each turbulence diagnostic are optimally combined to maximize the AUC. The data used have a forecast lead time between $T + 24$ and $T + 33$ hr between September 2016 and August 2017

4 | CONCLUSIONS

This study has expanded on the work by Storer *et al.* (2018a) by investigating the forecast skill of a multi-diagnostic multi-model ensemble for aviation turbulence. As in Storer *et al.* (2018a), probabilistic forecasts were created, but this time for five turbulence diagnostics (Ellrod and Knapp (1992) Turbulence Index 1, Brown Index, mountain wave turbulence, Richardson number and convective precipitation accumulation) and two ensembles: Met Office Global and Regional Ensemble Prediction System (MOGREPS-G) and European Centre for Medium-Range Weather Forecasts (ECMWF) Ensemble Prediction System (EPS). By combining the predictors, a multi-diagnostic predictor was created. Then, combining the predictors from the two ensembles, a multi-diagnostic multi-model ensemble was created. The trial ran from September 2016 to August 2017 and used the 0000 UTC forecast run and forecasts at $T + 24$, $T + 27$, $T + 30$ and $T + 33$ hr. The forecast was verified against the derived equivalent vertical gust derived from data from a fleet of Boeing 747 and 777 aircraft.

The results in this study agree with Storer *et al.* (2018a) that EPS51 is more skilful than the MOGREPS-G for the individual diagnostics. The multi-model ensemble was also more skilful than either of the single-model ensembles (but not significantly for most examples). When combining the predictors, the multi-diagnostic predictor was shown to be more skilful for MOGREPS-G than EPS51 for both the equal combined and optimized ensembles. Again, the multi-diagnostic multi-model ensemble is more skilful than the two single-model ensembles. The relative economic value plot of the multi-diagnostic shows that MOGREPS-G has greater value than EPS51, and in some cases also has greater value than the multi-model ensemble. All ensembles were shown to have similar reasonable reliability; however, above 1.25% forecast probability the ensembles start under-forecasting the events. To overcome this it could be possible to apply a nonlinear calibration; however, a linear calibration was kept in this study as it is suitable for most of the forecast probabilities.

The results from using a 12-member ECMWF ensemble (EPS12) show that EPS12 for the individual diagnostics is less skilful than EPS51 (as in Storer *et al.*, 2018a). When combined into the optimized multi-diagnostic ensemble, EPS12 was slightly more skilful. This therefore indicates that a smaller ensemble spread for the individual diagnostics within a multi-diagnostic ensemble is important for optimizing operationally in the future. Further studies with a larger sample would be useful to investigate this further. If EPS12 really does provide a more skilful forecast than EPS51 when used

in this way, it could reduce computation costs for turbulence forecasting. Only 12 members would need to be calculated, saving memory space and computational time. It is also worth exploring the impact of ensemble size on a multi-diagnostic multi-model ensemble further to see if there is an optimum number of ensemble members to maximize forecast skill.

MOGREPS-G is now running with 18 members and is designed to be time lagged to create a larger ensemble; therefore, investigating how well a 36-member MOGREPS-G performs against the 12-member MOGREPS-G would also be worth further study. Not only increasing the ensemble size but adding a third ensemble would also be worth investigating. The current operational World Area Forecast Centre forecasts are produced from a blend of both UK and US deterministic models and therefore the natural progression would be to add in the Global Ensemble Forecast System from the National Center for Environmental Prediction. This could provide more skill by adding its own strengths and weaknesses, but would need verifying before being made operational.

ACKNOWLEDGEMENTS

The authors would like to thank the two anonymous reviewers and editor for their helpful suggestions to improve this paper. The authors would also like to thank Ken Mylne and Piers Buchanan for helpful comments when reviewing this paper.

This work was carried out under Natural Environment Research Council, Grant/Award Number NE/L002566; Royal Society, Grant/Award Number UF130571.

ORCID

Luke N. Storer https://orcid.org/0000-0002-1805-4934
Philip G. Gill https://orcid.org/0000-0002-8884-3090
Paul D. Williams https://orcid.org/0000-0002-9713-9820

REFERENCES

Atlas, D., Metcalf, J., Richter, J. and Gossard, E. (1970) The birth of 'CAT' and microscale turbulence. *Journal of the Atmospheric Sciences*, 27(6), 903–913.

Brown, R. (1973) New indices to locate clear air turbulence. *Meteorological Magazine*, 102, 347–361.

Buizza, R. and Palmer, T.N. (1998) Impact of ensemble size on ensemble prediction. *Monthly Weather Review*, 126(9), 2503–2518.

Chambers, E. (1955) Clear air turbulence and civil jet operations. *Aeronautical Journal*, 59(537), 613–628.

Ehrendorfer, M. and Murphy, A.H. (1988) Comparative evaluation of weather forecasting systems: sufficiency, quality, and accuracy. *Monthly Weather Review*, 116(9), 1757–1770.

Ellrod, G.P. and Knapp, D.I. (1992) An objective clear-air turbulence forecasting technique: verification and operational use. *Weather and Forecasting*, 7(1), 150–165.

Endlich, R.M. (1964) The mesoscale structure of some regions of clear-air turbulence. *Journal of Applied Meteorology*, 3(3), 261–276.

Gill, P.G. (2014) Objective verification of World Area Forecast Centre clear air turbulence forecasts. *Meteorological Applications*, 21(1), 3–11.

Gill, P.G. (2016) Aviation turbulence forecast verification. In: Sharman, R., & Lane, T. (Eds.) *Aviation Turbulence*: Processes, Detection, Prediction. Switzerland: Springer International Publishing, pp. 261–283.

Gill, P.G. and Buchanan, P. (2014) An ensemble based turbulence forecasting system. *Meteorological Applications*, 21(1), 12–19.

Gill, P.G. and Stirling, A.J. (2013) Including convection in global turbulence forecasts. *Meteorological Applications*, 20(1), 107–114.

Gultepe, I., Sharman, R., Williams, P.D., Zhou, B., Ellrod, G., Minnis, P., Trier, S., Griffin, S., Yum, S.S., Gharabaghi, B., Feltz, W., Temimi, M., Pu, Z., Storer, L.N., Kneringer, P., Weston, M.J., Chuang, H.-Y.L., Thobois, L., Dimri, A.P., Dietz, S.J., França, G.B., Almeida, M.V. and Albquerque Neto, F.L. (2019) A review of high impact weather for aviation meteorology. *Pure and Applied Geophysics*, 176, 1869–1921.

ICAO. (2016) *Guidance on the harmonized WAFS grids for cumulonimbus cloud, icing and turbulence forecasts*. Available at: https://www.icao.int/safety/meteorology/WAFSOPSG/Guidance Material/Forms/AllItems.aspx [Accessed 20th March 2020].

Iris. (2015) *Met Office*. Available at: https://scitools.org.uk/iris/docs/latest/ [Accessed 20th March 2020].

Jolliffe, I.T. and Stephenson, D.B. (2012) *Forecast Verification: A Practitioner's Guide in Atmospheric Science*, 2nd edition. Chichester, UK: Wiley-Blackwell. John Wiley & Sons.

Kim, J.-H., Sharman, R., Strahan, M., Scheck, J., Bartholomew, C., Cheung, J., Buchanan, P. and Gait, N. (2018) Improvements in non-convective aviation turbulence prediction for the world area forecast system (WAFS). *Bulletin of the American Meteorological Society*, 99(11), 2295–2311.

Koch, S.E. and Dorian, P.B. (1988) A mesoscale gravity wave event observed during CCOPE. Part III: Wave environment and probable source mechanisms. *Monthly Weather Review*, 116(12), 2570–2592.

Lee, S.H., Williams, P.D. and Frame, T.H.A. (2019) Increased shear in the North Atlantic upper-level jet stream over the past four decades. *Nature*, 572, 639–642.

Lilly, D.K. (1978) A severe downslope windstorm and aircraft turbulence event induced by a mountain wave. *Journal of the Atmospheric Sciences*, 35(1), 59–77.

Owens, R.G. and Hewson, T.D. (2018) *ECMWF Forecast User Guide*. Reading: ECMWF. https://doi.org/10.21957/m1cs7h.

Richardson, D.S. (2000) Skill and relative economic value of the ECMWF ensemble prediction system. *Quarterly Journal of the Royal Meteorological Society*, 126(563), 649–667.

Sharman, R. and Lane, T. (2016) *Aviation Turbulence: Processes, Detection, Prediction*. Switzerland: Springer International Publishing.

Sharman, R.D. and Pearson, J.M. (2017) Prediction of energy dissipation rates for aviation turbulence. Part I: Forecasting

nonconvective turbulence. *Journal of Applied Meteorology and Climatology*, 56(2), 317–337.

Sharman, R., Tebaldi, C., Wiener, G. and Wolff, J. (2006) An integrated approach to mid- and upper-level turbulence forecasting. *Weather and Forecasting*, 21(3), 268–287.

Storer, L.N., Gill, P.G. and Williams, P.D. (2018a) Multi-model ensemble predictions of atmospheric turbulence. *Meteorological Applications*, 26(3), 416–428. https://doi.org/10.1002/met.1772.

Storer, L.N., Williams, P.D. and Gill, P.G. (2019) Aviation turbulence: dynamics, forecasting, and response to climate change. *Pure and Applied Geophysics*, 176, 2085–2091. https://doi.org/10.1007/s00024-018-1822-0.

Storer, L.N., Williams, P.D. and Joshi, M.M. (2017) Global response of clear-air turbulence to climate change. *Geophysical Research Letters*, 44(19), 9976–9984.

Truscott, B. (2000) *EUMETNET AMDAR AAA AMDAR Software Developments Technical Specification. Doc. Ref. E_AMDAR/TSC/003.* Exeter: Met Office.

Uccellini, L.W. and Koch, S.E. (1987) The synoptic setting and possible energy sources for mesoscale wave disturbances. *Monthly Weather Review*, 115(3), 721–729.

Williams, P.D. (2017) Increased light, moderate, and severe clear-air turbulence in response to climate change. *Advances in Atmospheric Sciences*, 34(5), 576–586.

Williams, P.D. and Joshi, M.M. (2013) Intensification of winter transatlantic aviation turbulence in response to climate change. *Nature Climate Change*, 3(7), 644–648.

How to cite this article: Storer LN, Gill PG, Williams PD. Multi-diagnostic multi-model ensemble forecasts of aviation turbulence. *Meteorol Appl*. 2020;27:e1885. https://doi.org/10.1002/met.1885